Canadian Hospital Re-admittance Challenge Project Report

Team : JUST BINARY SEARCH!
Adithya Sunilkumar (IMT2021068)
Kevin Adesara (IMT2021070)
Anant Ojha (IMT2021102)

October 23, 2023

Contents

1 Introduction

1.1 Background

The "Canadian Hospital Re-admittance Challenge" aims to reduce costly and disruptive hospital readmissions by applying data science and predictive modeling to identify at-risk patients. Analyzing a diverse dataset of patient information, this initiative seeks to enhance healthcare efficiency and patient outcomes through targeted interventions and improved discharge planning. It represents a crucial step in addressing the challenges of healthcare delivery and quality in Canada.

1.2 Objectives

1. Develop predictive models to identify high-risk patients for hospital readmission.
2. Improve healthcare efficiency by providing targeted interventions and personalized care plans.
3. Reduce hospital readmission rates and enhance patient outcomes in the Canadian healthcare system.

2 Data Preprocessing

2.1 Dataset Description

This dataset consists of patient information, including unique encounter and patient identifiers, demographic details (race, gender, age), medical history (diagnoses), medication prescriptions, and indicators of changes in medication dosage. It also includes information about the timing and types of healthcare encounters, such as outpatient and emergency visits, as well as the primary outcome variable "readmission_id" indicating the days to inpatient readmission.

2.2 Data Cleaning

During the data preprocessing phase, we identified several columns that were either unrelated to our analysis or had a significant number of null values. As a result, we made the decision to exclude the following columns from the dataset, because they consisted of more than 30% null values:

- weight
- medical_specialty
- payer_code
- max_glu_serum
- A1Cresult

Drug Columns: All the columns related to drugs provided little to no information about the target variable, so they were also dropped. Upon calculating the mode and its frequency for the drug columns, we noticed around 99% of entries would contain the same value. Later, we will describe how this data was modified into three features to make it more useful.

2.3 Data Encoding

As part of data preprocessing, we utilized Label Encoding to transform non-numeric categorical variables into numeric representations for improved compatibility with machine learning models. The following columns were encoded:

- **race**: Encoded the 'race' column to represent different racial categories as numeric values.
- **age**: Transformed age groups into numeric values for use in our analysis.
- **gender**: Encoded gender as numerical values, with 'male' and 'female' represented accordingly.
- **diabetesMed**: Converted 'diabetesMed' to a binary numeric representation, where 'yes' is represented as 1 and 'no' as 0.
- **change**: Transformed the 'change' column into a binary numeric variable, with 'change' as 1 and 'no change' as 0.
- **diag_1**: Encoded primary diagnosis codes ('diag_1') into numerical format.
- **diag_2**: Applied Label Encoding to secondary diagnosis codes ('diag_2') to enable numeric analysis.
- **diag_3**: Encoded additional secondary diagnosis codes ('diag_3') for numeric compatibility.

Label Encoding facilitated the integration of these categorical variables into our machine learning models, allowing us to perform quantitative analyses effectively.

2.4 Handling Missing Data

During data preprocessing, we implemented a strategy to handle missing values effectively. The following procedures were applied:

1. Rows with More Than 2 Null Values: Rows with more than two null values were identified and subsequently dropped from the dataset. This decision was made to ensure the retention of data points with a sufficient amount of information for analysis.
2. Replacement with Mode: For columns where null values were observed but did not exceed the threshold, the missing values were replaced with the mode of the respective column. Utilizing the mode helped maintain the integrity of the dataset while addressing missing data points in a meaningful way.

These steps ensured that the dataset remained suitable for subsequent analysis while minimizing the impact of missing data on our results. The initial row count of the dataset was 71,236. After the removal of rows with more than two null values, the row count was reduced to 71,225.

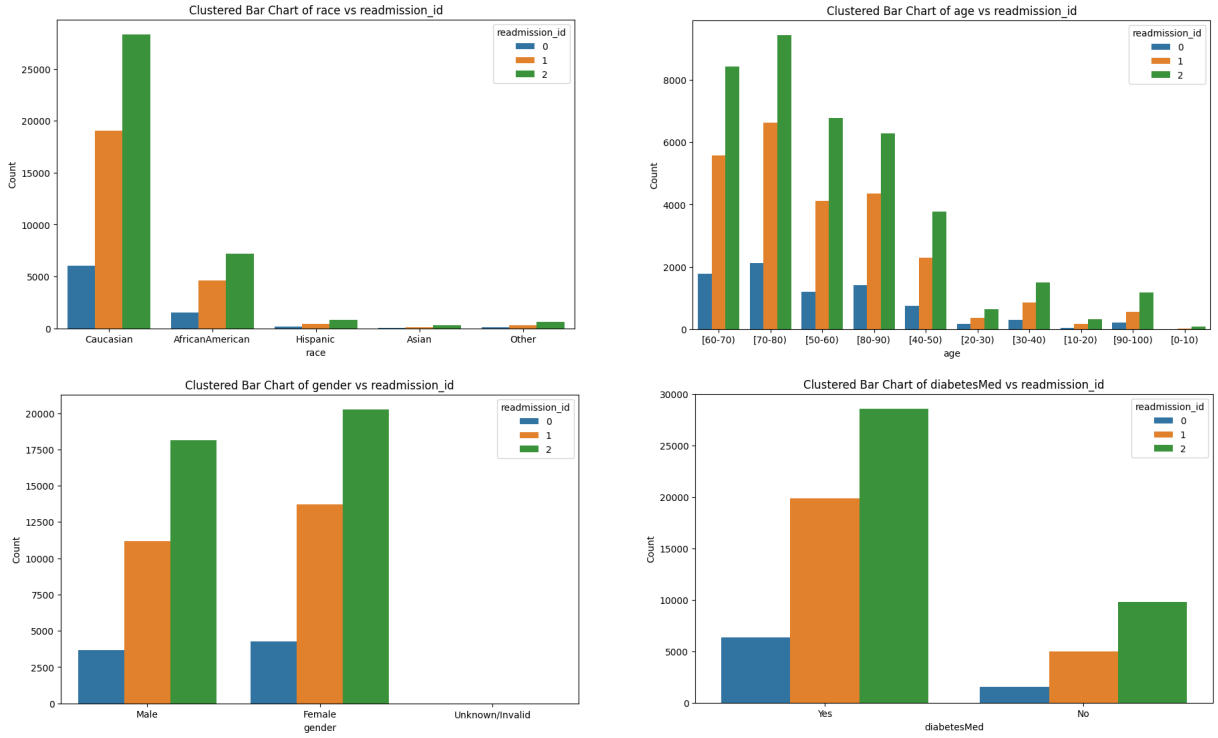
3 Exploratory Data Analysis (EDA)

3.1 Data Summary

The dataset contains a total of 71,236 records, each representing a patient encounter. It encompasses various features, including demographic details, medical history, medication information, and healthcare encounter attributes. The primary outcome variable, "readmission_id," categorizes patient encounters by days to inpatient readmission, with values 0, 1, and 2 indicating readmission within 30 days, beyond 30 days, and no recorded readmission, respectively.

3.2 Data Visualization

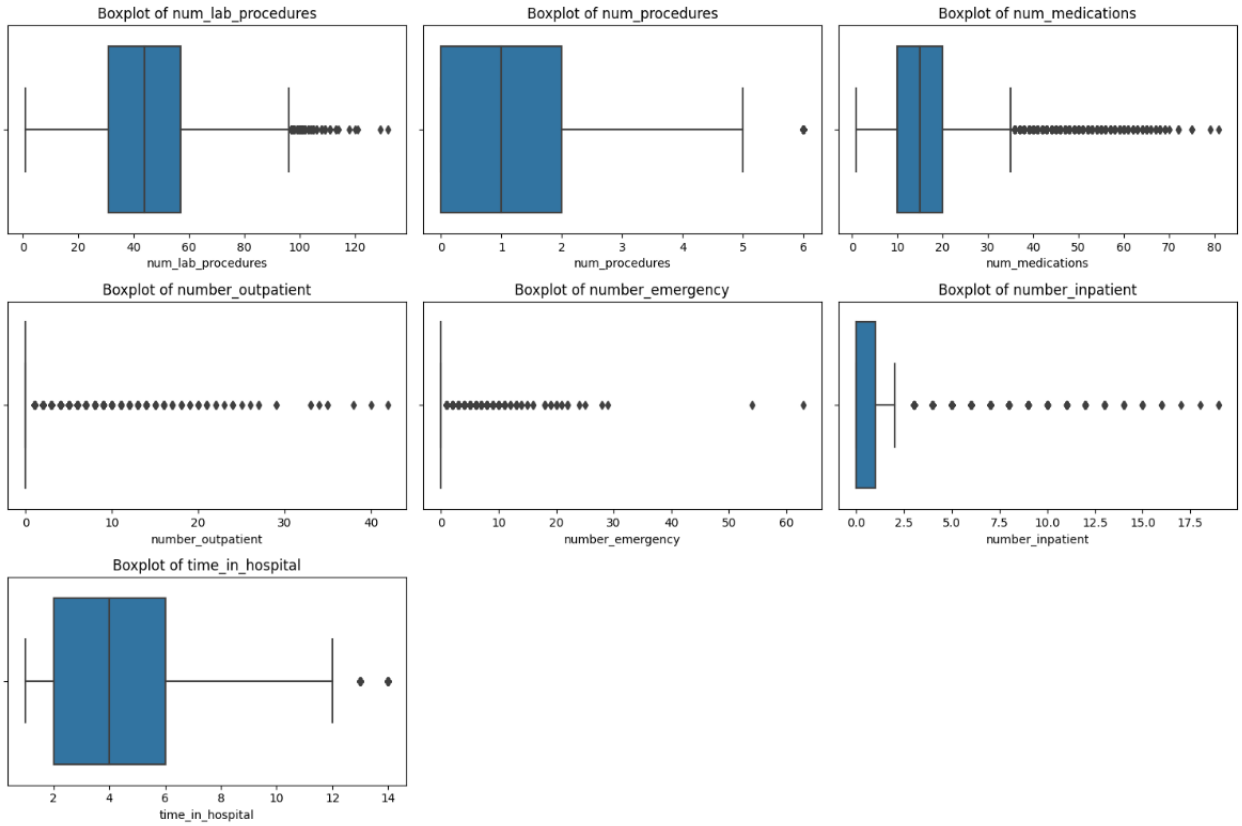
3.2.1 Clustered Bar Graph



We decided to treat the categorical and numeric features separately. For the former, we wanted to see the distribution of readmission_id by clustering the data belonging to a common category. Therefore, a bar graph was drawn for some of the categorical features as shown above. Note that age is also considered as a categorical feature here because it is given in ranges.

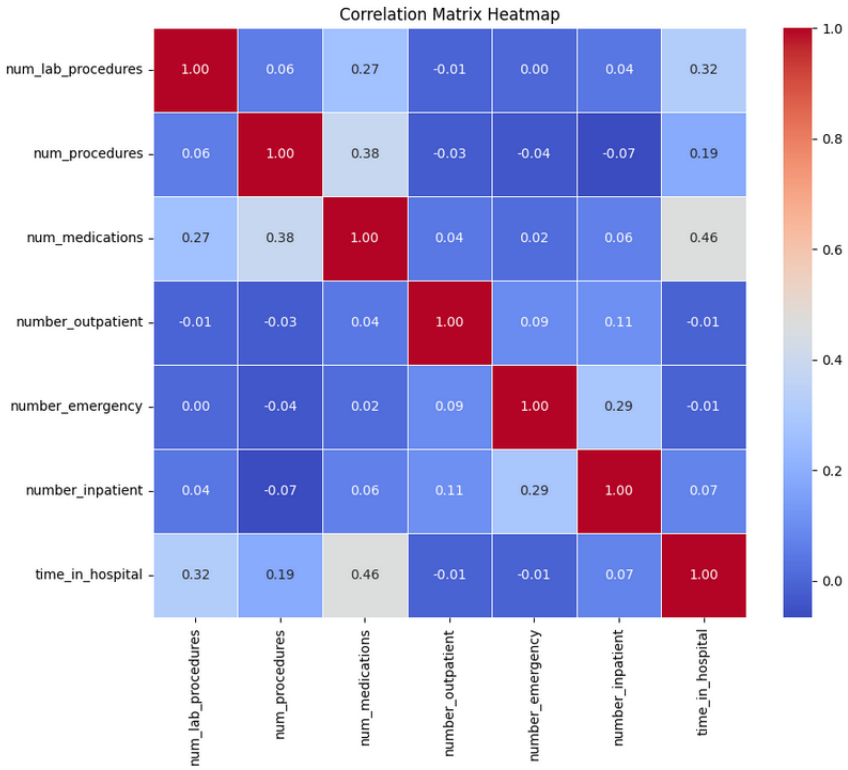
The colour of the bar represents the value of readmission_id and from the above graphs it was evident that irrespective of the feature or category, the readmission_id value 2 was most common followed by 1, and then 0. Unfortunately, there was no single category or feature which stood out, so it was logical to assume that one categorical feature alone did not play a large role in determining the readmission_id value.

3.2.2 Box Plots



For the numeric features, we have generated box plots, with whiskers extending 1.5 times the inter-quartile range. The above graphs give us some information about the data and the outliers. For example, number_outpatient and number_emergency columns have all positive values marked as outliers. Similarly, the number_inpatient column also has a very low upper-quartile and a median of 0.

3.2.3 Heatmap and Correlation Matrix



Again, for the numeric features, we calculated the correlation matrix and plotted the corresponding heatmap as shown above. In order for two variables to be moderately correlated, it should have a correlation coeffecient of above 0.5, but as shown in the figure, the highest correlation value is 0.46. This is between time in hospital and number of medications.

3.3 Post Analysis Actions

We introduced new columns to enrich the dataset and capture additional information:

- **Frequency of Patients**: A new column was created to record the frequency of patient encounters. This column provides insights into how often individual patients were encountered in the dataset, which can be valuable for certain analyses.
- **Frequency of Drug Changes (Up, Down, Steady)**: To better understand the usage patterns of different drugs, we introduced three new columns, one for each type of drug change (up, down, and steady). These columns represent the frequency of changes in drug dosage during patient encounters, allowing us to assess trends in medication adjustments over time.

These additional columns enhance the dataset's richness and provide valuable information for our analysis.

4 Model Building

4.1 Model Training

Models were trained using the following methods:

1. Random Forest with Randomized Grid Search.
2. XGBoost.
3. AdaBoost.
4. K-Nearest Neighbors (KNN).

4.2 Model Evaluation

The performance of each model was assessed, and the accuracy on the validation set is as follows:

- Random Forest: 0.7233
- XGBoost: 0.73
- AdaBoost: 0.70
- KNN: 0.51